Assignment Ten: Node Purity in Decision Trees

Brandon Trinkle

Arizona State University

Course Number: IFT 360

Professor Durgesh Sharma

9/21/24

Node Purity in Decision Trees

1. The calculation for entropy is $H(V) = -\sum_{k} P(v_k) log_2 P(v_k)$ (Artificial Intellligence: A Modern Approach, 2021, p. 661)

    a. Node 1:

        i. P(Sick) = 17/30

        ii. P(Healthy) = 13/30

$$H_1 = -\left(\frac{17}{30} log_2 \left(\frac{17}{30}\right) + \frac{13}{30} log_2 \left(\frac{13}{30}\right)\right) = 0.9871$$

    b. Node 2:

        i. P(Sick) = 12/30

        ii. P(Healthy) = 18/30

$$H_2 = -\left(\frac{12}{30} log_2 \left(\frac{12}{30}\right) + \frac{18}{30} log_2 \left(\frac{18}{30}\right)\right) = 0.9709$$

    c. Node 3:

        i. P(Sick) = 15/30

        ii. P(Healthy) = 15/30

$$H_3 = -\left(\frac{15}{30} log_2 \left(\frac{15}{30}\right) + \frac{15}{30} log_2 \left(\frac{15}{30}\right)\right) = 1.0$$

2. Node 2 is the purest as it has the lowest entropy value (0.9709), indicating the highest level of purity or the least uncertainty.

3. Node 3 is the most impure since it has the highest entropy value (1.0), indicating maximum uncertainty or impurity.

References

Artificial Intellligence: A Modern Approach. (2021). In S. Russell, & P. Norvig. Hoboken, NJ: Pearson.