Homework 2

Chonglin Zhang
833003072

March 4, 2024

Problem 1: Information Gain

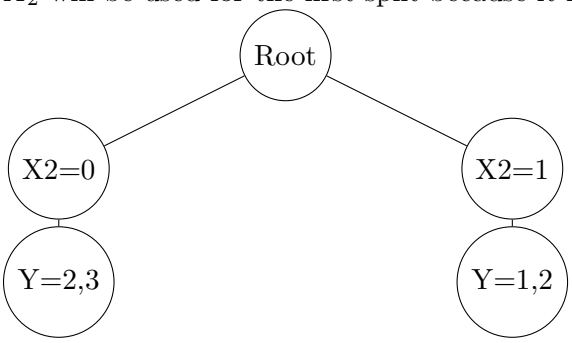
1. Calculate the conditional entropy for both $X_1 and X_2$.
H(Y | X_1) = - $[\frac{2}{6}(\frac{0}{2}\log(\frac{0}{2}) + \frac{1}{2}\log(\frac{1}{2}) + \frac{1}{2}\log(\frac{1}{2})) + \frac{4}{6}(\frac{2}{4}\log(\frac{2}{4}) + \frac{1}{2}\log(\frac{1}{2}) + \frac{1}{2}\log(\frac{1}{2}))] = -[-\frac{2}{6} - 1] = \frac{4}{3}$

H(Y | X_2) = - $[\frac{3}{6}(\frac{2}{3}\log(\frac{2}{3}) + \frac{1}{3}\log(\frac{1}{3}) + \frac{0}{3}\log(\frac{0}{3})) + \frac{3}{6}(\frac{0}{3}\log(\frac{3}{3}) + \frac{1}{3}\log(\frac{1}{3}) + \frac{2}{3}\log(\frac{2}{3}))] = 0.918$

2. Calculate the information gain if we split based on $X_1 or X_2$.

$H(Y) = 3 * \frac{1}{3}\log\left(\frac{1}{3}\right) = 1.585$
$IG(X_1) = 1.585 - 1.333 = 0.252$
$IG(X_2) = 1.585 - 0.918 = 0.667$

3. Report which attribute is used for the first split. Draw the decision tree using this split.
X_2 will be used for the first split because it has higher information gain than X_1



4. $X_1 = 0$ and $X_2 = 1$ which will be classified as Y=1 on this tree since X2=1.