Casey Pei
CSCE 421-500
Texas A&M University

March 9, 2024

HOMEWORK 2 — Tree-based Models

1 Math Questions

1.1 Information Gain (20 points)

NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

.

Suppose you are given 6 training points as seen below, for a classification problem with two binary attributes X_1 and X_2 and three classes $Y \in 1, 2, 3$. You will use a decision tree learner based on information gain

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

1. Calculate the conditional entropy for both X_1 and X_2.

   First, I will calculate the conditional probabilities for X_1 and X_2 on Y.

$P(X_1 \mid Y)$	$P(X_2 \mid Y)$
$P(X_1 = 1 \mid Y = 1) = \frac{2}{3}$	$P(X_2 = 1 \mid Y = 1) = \frac{3}{3}$
$P(X_1 = 0 \mid Y = 1) = \frac{1}{3}$	$P(X_2 = 0 \mid Y = 1) = \frac{0}{3}$
$P(X_1 = 1 \mid Y = 2) = \frac{1}{1}$	$P(X_2 = 1 \mid Y = 2) = \frac{0}{1}$
$P(X_1 = 0 \mid Y = 2) = \frac{0}{1}$	$P(X_2 = 0 \mid Y = 2) = \frac{1}{1}$
$P(X_1 = 1 \mid Y = 3) = \frac{0}{1}$	$P(X_2 = 1 \mid Y = 3) = \frac{0}{1}$
$P(X_1 = 0 \mid Y = 3) = \frac{1}{1}$	$P(X_2 = 0 \mid Y = 3) = \frac{1}{1}$

   Now with the conditional probabilities, we can calculate the conditional entropy of X_1 on Y.

$$H(X_1|Y) = -\sum P(X_1|Y) \times log_2(P(X_1|Y)) \tag{0-1}$$

$$= -\left[\frac{2}{3}log_2\left(\frac{2}{3}\right) + \frac{1}{3}log_2\left(\frac{1}{3}\right) + \frac{1}{1}log_2\left(\frac{1}{1}\right) + \frac{0}{1}log_2\left(\frac{0}{1}\right) + \frac{0}{1}log_2\left(\frac{0}{1}\right) + \frac{1}{1}log_2\left(\frac{1}{1}\right)\right] \tag{0-2}$$

$$= -\left[\frac{2}{3}(-0.585) + \frac{1}{3}(-1.585) + 0 + 0 + 0 + 0\right] \tag{0-3}$$

$$= 0.918 \tag{0-4}$$

   Now with the conditional probabilities, we can calculate the conditional entropy of X_2 on Y.

$$H(X_2|Y) = -\sum P(X_2|Y) \times log_2(P(X_2|Y)) \tag{0-5}$$

$$= -\left[\frac{3}{3}log_2\left(\frac{3}{3}\right) + \frac{0}{3}log_2\left(\frac{0}{3}\right) + \frac{0}{1}log_2\left(\frac{0}{1}\right) + \frac{1}{1}log_2\left(\frac{1}{1}\right) + \frac{0}{1}log_2\left(\frac{0}{1}\right) + \frac{1}{1}log_2\left(\frac{1}{1}\right)\right] \tag{0-6}$$

$$= -\left[\frac{3}{3}(0) + \frac{0}{3}(0) + 0 + 0 + 0 + 0\right] \tag{0-7}$$

$$= 0 \tag{0-8}$$

2. Calculate the information gain if we split based on 1) X_1 or 2) X_2

   To calculate the information gain for splitting based on X_1 or X_2, we first need to compute the entropy of the parent node and then the entropy of the child nodes after the split. Information gain is defined as the difference between the entropy of the parent node and the weighted average of the entropy of the child nodes.

   Let's start by calculating the entropy of the parent node:

   Number of instances for each class Y:

   - $Y = 1$: 2 instances
   - $Y = 2$: 1 instance
   - $Y = 3$: 3 instances
   - Total number of instances = 6

   Entropy of the parent node:

$$H(Y) = -\sum_{i=1}^{3} P(Y = i) \cdot \log_2(P(Y = i)) \tag{0-9}$$

$$H(Y) = -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) \tag{0-10}$$

$$H(Y) = -\left(\frac{1}{3} \cdot (-1.585) + \frac{1}{6} \cdot (-2.585) + \frac{1}{2} \cdot (-1)\right) \tag{0-11}$$

$$H(Y) = -\left(\frac{1}{3} \cdot (-1.585) + \frac{1}{6} \cdot (-2.585) + \frac{1}{2} \cdot (-1)\right) \tag{0-12}$$

$$H(Y) = -(-0.528 + -0.431 + -0.5) \tag{0-13}$$

$$H(Y) = 1.459 \tag{0-14}$$

   Now, let's calculate the information gain for splitting based on X_1 and X_2:

   For X_1:

   Child nodes after splitting based on X_1:

   - $X_1 = 1$: 4 instances (Y=1: 2, Y=2: 1, Y=3: 1)
   - $X_1 = 0$: 2 instances (Y=2: 1, Y=3: 1)

   Entropy of the child nodes after splitting based on X_1:

$$H(Y|X_1) = \sum_{j=1}^{2} P(X1 = j) \cdot H(Y|X_1 = j) \tag{0-15}$$

$$H(Y|X_1) = \left(\frac{4}{6} \cdot H(Y|X_1 = 1) + \frac{2}{6} \cdot H(Y|X_1 = 0)\right) \tag{0-16}$$

$$H(Y|X_1) = \left(\frac{4}{6} \cdot \text{Entropy}(Y|X_1 = 1) + \frac{2}{6} \cdot \text{Entropy}(Y|X_1 = 0)\right) \tag{0-17}$$

$$H(Y|X_1) = \left(\frac{4}{6} \cdot \left(-\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)\right) + \frac{2}{6} \cdot \left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right)\right) \tag{0-18}$$

$$H(Y|X_1) = \left(\frac{2}{3} \cdot (1.5) + \frac{1}{3} \cdot (1)\right) \tag{0-19}$$

$$H(Y|X_1) = \left(1 + \frac{1}{3}\right) \tag{0-20}$$

$$H(Y|X_1) = 1.333 \tag{0-21}$$

Information gain for splitting based on X_1:

$$IG(X_1) = H(Y) - H(Y|X_1) \tag{0-22}$$

$$IG(X_1) = 1.459 - 1.333 \tag{0-23}$$

$$IG(X_1) = 0.126 \tag{0-24}$$

For X_2:

Child nodes after splitting based on X_2:

- $X_2 = 1$: 3 instances (Y=1: 2, Y=2: 0, Y=3: 1)
- $X_2 = 0$: 3 instances (Y=1: 0, Y=2: 1, Y=3: 2)

Entropy of the child nodes after splitting based on X_2:

$$H(Y|X_2) = \sum_{j=1}^{2} P(X_2 = j) \cdot H(Y|X_2 = j) \tag{0-25}$$

$$H(Y|X_2) = \left(\frac{3}{6} \cdot H(Y|X_2 = 1) + \frac{3}{6} \cdot H(Y|X_2 = 0)\right) \tag{0-26}$$

$$H(Y|X_2) = \left(\frac{3}{6} \cdot \text{Entropy}(Y|X_2 = 1) + \frac{3}{6} \cdot \text{Entropy}(Y|X_2 = 0)\right) \tag{0-27}$$

$$H(Y|X_2) = \left(\frac{3}{6} \cdot \left(-\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)\right) + \frac{3}{6} \cdot \left(-\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)\right)\right) \tag{0-28}$$

$$H(Y|X_2) = \left(\frac{1}{2} \cdot (0.918) + \frac{1}{2} \cdot (0.918)\right) \tag{0-29}$$

$$H(Y|X_2) = (0.459 + 0.459) \tag{0-30}$$

$$H(Y|X2) = 0.918 \tag{0-31}$$

Information gain for splitting based on $X2$:

$$IG(X_2) = H(Y) - H(Y|X_2) \tag{0-32}$$

$$IG(X_2) = 1.459 - 0.918 \tag{0-33}$$

$$IG(X2) = 0.541 \tag{0-34}$$

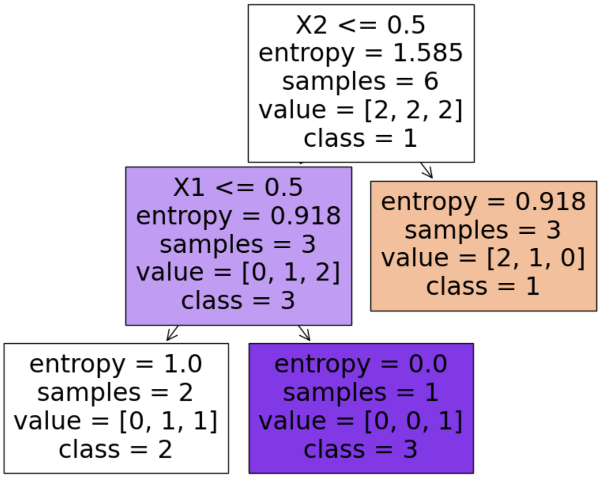
Therefore, the information gain for each is:

(a) X_1 is 0.126

(b) X_2 is 0.541

3. Report which attribute is used for the first split. Draw the decision tree using this split.

   Since the information gain for splitting based on X_2 (0.541) is greater than the information gain for splitting based on X_1 (0.126), the first split in the decision tree will be based on X_2.



4. Conduct classification for the test example $X_1 = 0$ and $X_2 = 1$.

   Following the above decision tree, we predict that $Y = 1$ given $X_1 = 0$ and $X_2 = 1$.

2 Programming Questions

Answers for these are located in the attached Casey_Pei_HW2.ipynb file.

Submitted by Casey Pei on March 9, 2024.