CSCE 633 : Machine Learning, Assignment - 2

Neha Joshi

8 March 2024

Question 1

Suppose you are given 6 training points for a classification problem with two binary attributes $X1$ and $X2$ and three classes $Y \in \{1, 2, 3\}$. You will use a decision tree learner based on information gain.

$X1$	$X2$	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

Answer

(1) Conditional Entropy for $X1$ and $X2$

Conditional entropy $H(Y|X)$ is given by:

$$H(Y|X) = -\sum_{x \in X} P(X = x) \sum_{y \in Y} P(Y = y|X = x) \log_2 P(Y = y|X = x)$$

For attribute $X1$:

$$P(X1 = 1) = \frac{4}{6} = 0.66$$

$$P(X1 = 0) = \frac{2}{6} = 0.33$$

For $X1 = 1$:

$$P(Y = 1|X1 = 1) = \frac{2}{4} = 0.5$$

$$P(Y = 2|X1 = 1) = \frac{1}{4} = 0.25$$

$$P(Y = 3|X1 = 1) = \frac{1}{4} = 0.25$$

For $X1 = 0$:

$$P(Y = 1|X1 = 0) = 0$$

$$P(Y = 2|X1 = 0) = \frac{1}{2} = 0.5$$

$$P(Y = 3|X1 = 0) = \frac{1}{2} = 0.5$$

So, for $X1$:

$$H(Y|X1) = -[0.66 \cdot (0.5 \cdot \log_2 0.5 + 0.25 \cdot \log_2 0.25 + 0.25 \cdot \log_2 0.25) + 0.33 \cdot (0 + 0.5 \cdot \log_2 0.5 + 0.5 \cdot \log_2 0.5)] = \mathbf{1.32}$$

For attribute $X2$:

$$P(X2 = 1) = \frac{3}{6} = 0.5$$

$$P(X2 = 0) = \frac{3}{6} = 0.5$$

For $X2 = 1$:

$$P(Y = 1|X2 = 1) = \frac{2}{3} = 0.66$$

$$P(Y = 2|X2 = 1) = \frac{1}{3} = 0.33$$

$$P(Y = 3|X2 = 1) = 0$$

For $X2 = 0$:

$$P(Y = 1|X2 = 0) = 0$$

$$P(Y = 2|X2 = 0) = \frac{1}{3} = 0.33$$

$$P(Y = 3|X2 = 0) = \frac{2}{3} = 0.66$$

So, for $X2$:

$$H(Y|X2) = -[(0.5 \cdot (0.66 \cdot \log_2 0.66 + 0.33 \cdot \log_2 0.33 + 0) + (0.5 \cdot (0.66 \cdot \log_2 0.66 + 0.33 \cdot \log_2 0.33 + 0)] = \mathbf{0.9234}$$

(2) Information Gain

Information gain is given by:

$$InfoGain(Y, X) = H(Y) - H(Y|X)$$

So,

$$InfoGain(Y, X1) = H(Y) - H(Y|X1)$$

$$InfoGain(Y, X2) = H(Y) - H(Y|X2)$$

$$H(Y) = -\left(\sum_{i=1}^{n} P(Y = i)\log_2 P(Y = i)\right) = -[((0.33 \cdot \log_2 0.33 + 0.33 \cdot \log_2 0.33 + 00.33 \cdot \log_2 0.33))] = 1.583$$

So,

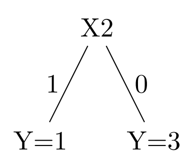
$$InfoGain(Y, X1) = H(Y) - H(Y|X1) = 1.583 - 1.32 = \mathbf{0.2633}$$

$$InfoGain(Y, X2) = H(Y) - H(Y|X2) = 1.583 - 0.9234 = \mathbf{0.6596}$$

(3) Decision Tree

The attribute with the highest information gain will be used for the first split.

Since information gain of X2 is greater than X1, we start with X2



(4) Classification for Test Example

From the tree above, since the decision tree splits based on $X2$ classified as $\mathbf{Y = 1}$.