Homework 2

Mu-Ruei Tseng

March 9, 2024

1 Problem 1: Information Gain (20 points)

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.

1. Calculate the conditional entropy for both X_1 and X_2.
2. Calculate the information gain if we split based on 1) X_1 or 2) X_2.
3. Report which attribute is used for the first split. Draw the decision tree using this split.
4. Conduct classification for the test example $X_1 = 0$ and $X_2 = 1$.

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

1.1 Solution

1.1.1 Calculate the conditional entropy for both X_1 and X_2

The definition of Conditional Entropy is:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \tag{1}$$

and

$$H(Y|X = x) = -\sum_{y \in \mathcal{Y}} p(y|X = x) \log_2(p(y|X = x)) \tag{2}$$

Conditional Entropy for X_1

$$H(Y|X_1) = \sum_{x \in \{0,1\}} p(x) H(Y|X_1 = x) \tag{3}$$

Since we have:

$$H(Y|X_1=0) = -\sum_{y\in\{1,2,3\}} p(y|X_1=0)\log_2(p(y|X_1=0)) = -(\frac{0}{2}\log_2\frac{0}{2} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1 \tag{4}$$

$$H(Y|X_1=1) = -\sum_{y\in\{1,2,3\}} p(y|X_1=1)\log_2(p(y|X_1=1)) = -(\frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}) = 1.5 \tag{5}$$

Therefore the conditional entropy for X_1 is:

$$H(Y|X_1) = \frac{2}{6}\times 1 + \frac{4}{6}\times 1.5 = \frac{4}{3} \tag{6}$$

Conditional Entropy for X_2

$$H(Y|X_2) = \sum_{x\in\{0,1\}} p(x)H(Y|X_2=x) \tag{7}$$

Since we have:

$$H(Y|X_2=0) = -\sum_{y\in\{1,2,3\}} p(y|X_2=0)\log_2(p(y|X_2=0)) = -(\frac{0}{3}\log_2\frac{0}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) \approx 0.918 \tag{8}$$

$$H(Y|X_2=1) = -\sum_{y\in\{1,2,3\}} p(y|X_2=1)\log_2(p(y|X_2=1)) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{0}{3}\log_2\frac{0}{3}) \approx 0.918 \tag{9}$$

Therefore the conditional entropy for X_1 is:

$$H(Y|X_2) = \frac{3}{6}\times 0.918 + \frac{3}{6}\times 0.918 = 0.918 \tag{10}$$

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

The definition of Information Gain is:

$$IG(Y,X) = E(Y) - E(Y|X) \tag{11}$$

Since $E(Y)$ is:

$$E(Y) = -(\frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3}) \approx 1.585 \tag{12}$$

The information gain if we split based on

1. X_1: $E(Y) - E(Y|X_1) \approx 1.585 - 1.333 = 0.252$
2. X_2: $E(Y) - E(Y|X_2) \approx 1.585 - 0.918 = 0.667$

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

Since splitting with X_2 achieves a higher information gain, we use it for the first split. Therefore, our decision tree will look like:

Split on X_2

$X_2 = 0$ $\quad$ $X_2 = 1$

$Y = 3$ $\quad$ $Y = 1$

$((y_1, y_2, y_3) = (0, 1, 2))$ $\quad$ $((y_1, y_2, y_3) = (2, 1, 0))$

, where y_i is the number of data that has the label $Y = i$ after the split. The majority label within the split is used for assigning predicted labels.

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

According to the tree we formulated in the previous section, instances with $X_2 = 1$ are classified as class $Y = 1$. Therefore, the test example would be classified as class $Y = 1$.