CS-559 Machine Learning

Homework - 3

Problem 1: Decision Trees based on Entropy

(1) The Entropy for a given node t is given by -

$$\text{Entropy}(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

→ The overall entropy before splitting is,

$$\text{Entropy}(t) = -\sum_j p(j|t) \cdot \log_2 p(j|t)$$

$$= -\left[\frac{4}{9}\log_2\left(\frac{4}{9}\right) + \frac{5}{9}\log_2\left(\frac{5}{9}\right)\right]$$

$$= -\left[(-0.5199) + (-0.4711)\right]$$

$$= 0.9910$$

→ The contingency tables after splitting on attribute a_1,

	$a_1 = T$	$a_2 = F$
+	3	1
−	1	4

→ Entropy for attribute a_1,

$$E_{a_1=T} = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$$

$$= -\left[(-0.3112) + (-0.5)\right]$$

$$= 0.8112$$

$$E_{a_1=F} = -\left[\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right]$$

$$= -\left[(-0.4643) + (-0.2575)\right]$$

$$= 0.7218$$

⇒ Calculating the information gain,

$$\Delta_{a_1} = \text{Entropy}(t) - \left[\frac{4}{9} \times \text{Entropy}(a_1=T) + \frac{5}{9} \times \text{Entropy}(a_1=F)\right]$$

$$= 0.9910 - \left[4/9 \times (0.8112) + 5/9\,(0.7218)\right]$$

$$= 0.9910 - \left[0.7615\right]$$

$$\boxed{\Delta_{a_1} = 0.22946}$$

⇒ Since a_2 is a continuous attribute, the contingency table after splitting on $a_{2,1}$

Sorted values	1	3	4	5	6	7	8

Split position	0.5		2		3.5		4.5		5.5		6.5		7.5		8.5	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
−	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
Gain	0		0.1427		0.00248		0.0728		0.00722		0.0183		0.1021		0	

Now

⇒ split 1:

$$H(1) = -\sum_{i=1}^{c} p(j|t) \log_2 p(j|t)$$

$$= -\left[\frac{4}{9}\log_2\left(\frac{4}{9}\right) + \frac{5}{9}\log_2\left(\frac{5}{9}\right)\right]$$

$$= 0.99107$$

Information Gain = $E(t) - H(1)$

$= 0.9910 - 0.9910$

$= \underline{0}$

⇒ split 2:

$$H(2: <=) = -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + 0\log_2 0\right)$$

$= 0$

$$H(2: >) = -\left(\frac{3}{8}\log_2\left(\frac{3}{8}\right) + \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right)$$

$$= -\left[(-0.5306) + (-0.4237)\right]$$

$= 0.9543$

Weighted Average = $\left[\frac{1}{9} \times 0 + \frac{8}{9} \times 0.9543\right]$

$= 0.8482$

Information Gain = $0.9910 - 0.8482$

$= \underline{0.1427}$

⇒ split 3:

$$H(3.5: <=) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

$$= -\left[(-0.5) + (-0.5)\right] = 1$$

$$H(3.5: >) = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

$$= -\left[(-0.5238) + (-0.4613)\right]$$

$$= 0.9851$$

$$\text{Weighted Average} = \left(\frac{2}{9} \times 1 + \frac{7}{9} \times 0.98523\right)$$

$$= 0.98851$$

$$\text{Information Gain} = 0.9910 - 0.98851$$

$$= 0.00248$$

→ Split 4 :-

$$H(4.5: <=) = -\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$

$$= -\left[(-0.3899) + (-0.5283)\right]$$

$$= 0.9182$$

$$H(4.5: >) = -\left[\frac{2}{6}\log_2\left(\frac{2}{6}\right) + \frac{4}{6}\log_2\left(\frac{4}{6}\right)\right]$$

$$= -\left[(-0.5283) + (-0.3899)\right]$$

$$= 0.9182$$

$$\text{Weighted Average} = \left(\frac{3}{9} \times 0.9182 + \frac{6}{9} \times 0.9182\right)$$

$$= 0.9182$$

$$\text{Information Gain} = A = 0.9910 - 0.9182 = 0.0728$$

→ split 5:

$$H(5.5 : <=) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

$$= -\left[(-0.5287) + (-0.4421)\right]$$

$$= 0.9708$$

$$H(5.5 : >) = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$

$$= 1$$

$$\text{Weighted Average} = \left[\frac{5}{9} \times 0.9708 + 1 \times \frac{4}{9}\right]$$

$$= 0.9837$$

$$\text{Information Gain} = 0.9910 - 0.9837$$

$$= 0.00722$$

→ split 6:

$$H(6.5 : <=) = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

$$= 1$$

$$H(6.5 : >) = -\left[\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log\left(\frac{2}{3}\right)\right]$$

$$= -\left[(-0.5283) + (-0.38997)\right]$$

$$= 0.9182$$

$$\text{Weighted average} = \left[\frac{6}{9} \times 1 + \frac{3}{9} \times 0.9188\right] = 0.9792$$

Information Gain = 0.9910 − 0.9727

= 0.0183

→ Split 7:

$$H(7.5:<=) = -\left[\frac{4}{8}\log_2\left(\frac{4}{8}\right) + \frac{4}{8}\log_2\left(\frac{4}{8}\right)\right] = 1$$

$$H(7.5:>) = -\left[0\log_2 0 + 1\log_2 1\right] = 0$$

weighted average = $\frac{8}{9} \times 1 = 0.8889$

Information gain = 0.9910 − 0.8889

= 0.1021

→ Split 8:

$$H(8.5:<=) = -\left[\frac{4}{9}\log_2\left(\frac{4}{9}\right) + \frac{5}{9}\log_2\left(\frac{5}{9}\right)\right]$$

$$= -\left[(-0.5199) + (-0.4711)\right]$$

$$= 0.9910$$

$$H(8.5:>) = -\left[0\log_2(0) + 0\log_2(0)\right] = 0$$

Information gain = 0.9910 − 0.9910 = 0

⇒ For attribute a_2 the best split is 2.0 with information gain 0.1427

⇒ But attribute a_1 has higher information gain

⇒ Hence, best attribute for the first splitting is attribute a_1.

(2) If we use "Instance" as another attribute and split point = 2.5, we can get best split point for instance

Count matrix < 2.5

	<=	>
+	2	2
−	0	5

$$\text{Entropy}_{(2.5:<=)} = -\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right) + 0\log_2 0\right] = 0$$

$$\text{Entropy}\,(2.5:>) = -\left[\frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{5}{7}\log_2\left(\frac{5}{7}\right) = 0.86312\right.$$

$$\text{Information Gain} = H(t) - \frac{7}{9} \times \text{Entropy}$$

$$= 0.99107 - 0.6713$$

$$= 0.3198$$

By Comparing both attribute, instances has high ~~attribute~~ Information gain. Therefore instance as an attribute is a root node and it can be used an an attribute for a decision in a tree.