Homework Assignment 3

* Problem 1 : Decision Trees based on Entropy.

Instance	Attribute 1 (a_1)	Attribute 2 (a_2)	Class
1	T	1.0	+
2	T	6.0	+
3	T	5.0	-
4	F	4.0	+
5	F	7.0	-
6	F	3.0	-
7	F	8.0	-
8	T	7.0	+
9	F	5.0	-

(1)

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

* $\text{Gain}_{\text{split}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i)$ ...(ii)

Information Gain

Entropy (p): parent node

→ $P(+) = 4/9$ and $P(-) = 5/9$

$\therefore \text{Entropy}(p) = -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$

$= -\frac{4}{9}(-1.1699) - \frac{5}{9}(-0.8480)$

$= 0.5199 + 0.4711$

$= 0.9911$

* For attribute 1 (a_1), count matrix is,

a_1	+	-
T	3	1
F	1	4

→ Entropy $(T_0) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= -\frac{3}{4}(-0.415) - \frac{1}{4}(-2)$

$= 0.31125 + 0.5$

$= 0.81125$

→ Entropy $(F) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)$

$= -\frac{1}{5}(-2.322) - \frac{4}{5}(-0.322)$

$= 0.4644 + 0.2576$

$= 0.722$

⇒ Information gain for a_1,

$= 0.9911 - \left[\frac{4}{9}(0.81125) + \frac{5}{9}(0.722)\right]$

$= 0.9911 - [0.361 + 0.401]$

$= \boxed{0.2294}$ ... [from (ii)]

Class	+	−	+	−	−	+	+	−	−

Attribute 2 (a_2)

Sorted values	1.0	3.0	4.0	5.0	5.0	6.0	7.0	7.0	8.0

Splits	1.0		3.0		4.0		5.0		6.0		7.0		8.0	
	<	≥	<	≥	<	≥	<	≥	<	≥	<	≥	<	≥
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0
−	0	5	0	5	1	4	1	4	3	2	3	2	4	1

Info. Gain	IG(0.5)	IG(2.0)	IG(3.5)	IG(4.5)	IG(5.5)	IG(6.5)	IG(7.5)
=	−	0.1427	0.0026	0.0728	0.0072	0.0183	0.1022

→ The information gain at split 1.0 is not helpful, as it will give parent entropy.

(i) Entropy (3.0) $= \frac{1}{9}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right]$

$+ \frac{8}{9}\left[-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right]$

$= \frac{1}{9}(0) + \frac{8}{9}[0.5306 + 0.4238]$

$= 0.8484$

$\Rightarrow IG(3.0) = 0.9911 - 0.8484 = 0.1427$

(ii) Entropy (4.0) $= \frac{2}{9}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

$+ \frac{7}{9}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$

$= 2/9\,(0.5+0.5) + 7/9\,(0.5239 + 0.4613)$

$= 0.2222 + 0.7663$

$= 0.9885$

$\Rightarrow IG(4.0) = 0.9911 - 0.9885 = 0.0026$

(iii) Entropy (5.0) $= \frac{3}{9}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$

$+ \frac{6}{9}\left[-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)\right]$

$= 3/9(0.3900 + 0.5283)$

$+ \frac{2}{3}(0.5283 + 0.3900)$

$= 0.3061 + 0.6122$

$= 0.9183$

→ IG (5.0) $= 0.9911 - 0.9183 = 0.0728$

(iv) Entropy (6.0) $= \frac{5}{9}\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$

$+ \frac{4}{9}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$

$= 5/9(0.529 + 0.4422)$

$+ 4/9(0.5 + 0.5)$

$= 0.5396 + 0.4444$

$= 0.9839$

→ IG (6.0) $= 0.9911 - 0.9839 = 0.0072$

(v) Entropy (7.0) $= \frac{6}{9}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$

$+ \frac{3}{9}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]$

$= 2/3(0.5 + 0.5) + 1/3(0.5283 + 0.39)$

$= 0.6667 + 0.3061$

$= 0.9728$

→ IG (7.0) $= 0.9911 - 0.9728 = 0.0183$

(vi) Entropy (8.0) $= \frac{8}{9}\left[-\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right)\right]$

$+ \frac{1}{9}\left[-\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right]$

$= \frac{8}{9}[0.5 + 0.5] + 0$

$= 0.8889.$

$\rightarrow$ IG (8.0) $= 0.9911 - 0.8889 = 0.1022$

$\Rightarrow$ So, maximum information gain is at split = 3.0 which is $\boxed{0.1427}$ for attribute 2 (a_2)

$\Rightarrow$ Among a_1 and a_2, $\boxed{a_1}$ will be chosen as the first splitting for decision tree, as it has maximum information gain.

(2) I don't think that "Instance" attribute should be used for a decision in the tree, as for each new Instance, new number will be assigned and hence it has no predictive power.