Chapter 8. Ex 8

解：

(a) 为便于观察，现将样本数据重新组织如下：

a_1	a_2	a_3	a_4	类别
0	1	1	0	w_1
1	0	1	0	w_1
0	0	1	1	w_1
1	1	1	1	w_1
1	0	1	1	w_2
0	0	0	0	w_2
0	1	0	0	w_2
1	1	1	0	w_2

a_1、a_2、a_3、a_4 可看作每一位二进制数的特征。

则经验熵为 $H(D) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

1) 计算 a_1 熵增益

$$\begin{aligned} \Delta i(H)_{a_1} &= H(D) - H(D|a_1) \\ &= 1 - \left(\frac{1}{2}H(D_0) + \frac{1}{2}H(D_1)\right) \\ &= 1 - \left[\frac{1}{2}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{2}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)\right] \\ &= 0 \end{aligned}$$

其中 D_0 为 A 值为 0 的数据子集，
D_1 为 A 值为 1 的数据子集。

(2) 计算 a_2 的熵增益

$$\Delta i(N)a_2 = H(D) - H(D|a_2)$$
$$= 1 - \left[\frac{5}{8}H(D_0) + \frac{3}{8}H(D_1)\right]$$
$$= 0$$

(3) 计算 a_3 的熵增益

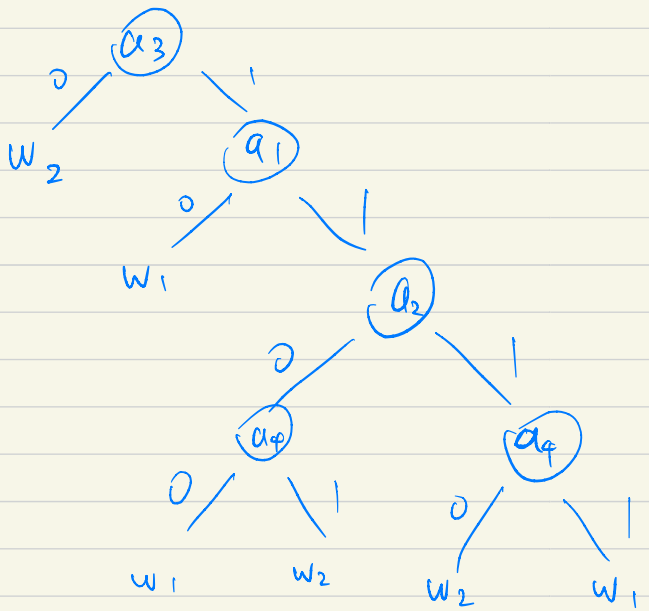
$$\Delta i(N)a_3 = H(D) - H(D|a_3)$$
$$= 1 - \left[\frac{2}{8}H(D_0) + \frac{6}{8}H(D_1)\right]$$
$$= 1 - \left[\frac{2}{8}\times 0 + \frac{6}{8}\cdot\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)\right]$$
$$\approx 1 - [0 + 0.6887]$$
$$= 0.31274$$

(4) 计算 a_4 的熵增益

$$\Delta i(N)a_4 = 1 - \left[\frac{5}{8}H(D_0) + \frac{3}{8}H(D_1)\right]$$
$$= 1 - \left[\frac{5}{8}\cdot\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{3}{8}\cdot\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)\right]$$
$$= 1 - \left[\frac{5}{8}\times(0.52877 + 0.42196) + \frac{3}{8}\times(0.38997 + 0.52832)\right]$$
$$= 0.0615$$

∵ $\Delta i(N)_{a_3}$ 最大，故第一个分支点选为 a_3

类似计算，可得判定树如下：



b).

$w_1 = (a_3 \wedge \neg a_1) \vee (a_3 \wedge a_1 \wedge \neg a_2 \wedge \neg a_4)$
$\vee (a_3 \wedge a_1 \wedge a_2 \wedge a_4)$

$w_2 = \neg a_3 \vee (a_3 \wedge a_1 \wedge \neg a_2 \wedge a_4) \vee (a_3 \wedge a_1 \wedge a_2 \wedge \neg a_4)$