

HW2 参考答案

一、解答：

(1) 划分前，母节点的信息熵为（如下对数符号都是以 2 为底数）：

$$E_{\text{parent}} = -0.4 \times \log 0.4 - 0.6 \times \log 0.6 = 0.9710$$

如果按照 Feature A 进行划分，我们可以得到：

	Feature A=T	Feature A=F
C0	4	0
C1	3	3

因此， $E_{A=T} = -3/7 \times \log(3/7) - 4/7 \times \log(4/7) = 0.9852$

$$E_{A=F} = -3/3 \times \log(3/3) - 0/3 \times \log(0/3) = 0$$

$$E_A = 7/10 \times 0.9852 + 3/10 \times 0$$

$$\Delta E_A = 0.9710 - E_A = \mathbf{0.2813}$$

同理，如果按照 Feature B 进行划分时，我们可以得到：

	Feature B=T	Feature B=F
C0	3	1
C1	1	5

因此， $E_{B=T} = -3/4 \times \log(3/4) - 1/4 \times \log(1/4) = 0.8113$

$$E_{B=F} = -1/6 \times \log(1/6) - 5/6 \times \log(5/6) = 0.65$$

$$E_B = 4/10 \times 0.8113 + 6/10 \times 0.65$$

$$\Delta E_B = 0.9710 - E_B = \mathbf{0.2565}$$

(2) $Gini_{\text{parent}} = 1 - 0.4 \times 0.4 - 0.6 \times 0.6 = 0.48$

$$Gini_{A=T} = 1 - (3/7) \times (3/7) - (4/7) \times (4/7) = 0.4898$$

$$Gini_{A=F} = 1 - 1 - 0 = 0$$

$$Gini_A = 7/10 \times 0.4898 + 3/10 \times 0$$

$$\Delta Gini_A = 0.48 - Gini_A = \mathbf{0.1371}$$

同理，可以算出： $\Delta Gini_B = 0.48 - Gini_B = \mathbf{0.1633}$

因为， $\Delta Gini_B > \Delta Gini_A$

所以，选择 Feature B 进行划分；

二、解答：

(1) 根据 Apriori 算法，写出详细过程：

频繁的 1-项集有：A (100%), B (75%), C (75%), E(50%), H(50%);

频繁的 2-项集有：AB (75%), AC(75%), AE(50%), AH(50%), BC(50%), BE(50%)

频繁的 3-项集有：ABC, ABE

(2) 根据定义，候选产生的宽度为 3 的强规则及其置信度为：

对于 ABC 而言，AB->C (2/3), AC->B (2/3), B->AC(2/3), C->AB(2/3), BC->A (1)

对于 ABE 而言，B->AE (2/3), E->AB (1), AB->E(2/3), AE->B(1), BE->A (1)