习题3

4、(1)

$$Entropy(S) = -\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right) \approx 0.971$$

① $$Entropy(S_{A=T}) = -\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right) \approx 0.9852$$

$$Entropy(S_{A=F}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$Entropy_A(S) = \frac{7}{10} \times Entropy(S_{A=T}) + \frac{3}{10} \times Entropy(S_{A=F})$$

$$= \frac{7}{10} \times 0.9852 + \frac{3}{10} \times 0 \approx 0.6896$$

$$Gain(S, A) = Entropy(S) - Entropy_A(S)$$

$$= 0.971 - 0.6896 = 0.2813$$

② $$Entropy(S_{B=T}) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \approx 0.8113$$

$$Entropy(S_{B=F}) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right) \approx 0.65$$

$$Entropy_B(S) = \frac{4}{10} \times 0.8113 + \frac{6}{10} \times 0.65 = 0.7145$$

$$Gain(S, B) = Entropy(S) - Entropy_B(S)$$

$$= 0.971 - 0.7145 = 0.2565$$

因为 $Gain(S,A) > Gain(S,B)$，所以算法将选择属性A.

(2) $$G(S) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48$$

① $$G(S_{A=T}) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898$$

$$G(S_{A=F}) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

[redacted] $$\Delta G(S,A) = G(S) - \left[\frac{7}{10}G(S_{A=T}) + \frac{3}{10}G(S_{A=F})\right] = 0.48 - \left[\frac{7}{10} \times 0.4898 + \frac{3}{10} \times 0\right] = 0.1371$$

② $G(S_{B=T}) = 1-(\frac{1}{4})^2-(\frac{3}{4})^2 = 0.375$

$G(S_{B=F}) = 1-(\frac{1}{6})^2-(\frac{5}{6})^2 = 0.2778$

$\Delta G(S,B) = G(S) - \left[\frac{4}{10}G(S_{B=T}) + \frac{6}{10}G(S_{B=F})\right]$

$= 0.48 - (\frac{4}{10}\times 0.375 + \frac{6}{10}\times 0.2778)$

$= 0.1633$

因为$\Delta G(S,B) > \Delta G(S,A)$，所以算法将选择属性B划分.

7. (1) $P(A|+) = \frac{3}{5}$ $P(A|-) = \frac{2}{5}$

$P(B|+) = \frac{1}{5}$ $P(B|-) = \frac{2}{5}$

$P(C|+) = \frac{4}{5}$ $P(C|-) = \frac{5}{5} = 1$

(2) ① $P(+|X) = P(X|+)\cdot P(+) \div P(X)$

记 $P_+ = P(X|+)\cdot P(+)$

$= P(A=0|+)\cdot P(B=1|+)\cdot P(C=0|+)\cdot P(+)$

$= (1-\frac{3}{5})\times \frac{1}{5}\times(1-\frac{4}{5})\times\frac{5}{10} = 0.008$

$P(-|X) = P(X|-)\cdot P(-) \div P(X)$

$P_- = P(X|-)\cdot P(-)$

$= P(A=0|-)\cdot P(B=1|-)\cdot P(C=0|-)\cdot P(-)$

$= (1-\frac{2}{5})\cdot\frac{2}{5}\cdot(1-\frac{5}{5})\cdot\frac{5}{10} = 0$

$P_+ > P_-$，所以此样本X预测类标号为+.

(3) $P(A|+) = \dfrac{3+m\cdot p}{5+m} = \dfrac{3+4\times\frac{1}{2}}{5+4} = \dfrac{5}{9}$

$P(A|-) = \dfrac{2+m\cdot p}{5+m} = \dfrac{2+2}{5+4} = \dfrac{4}{9}$

同理，$P(B|+) = \dfrac{1+2}{5+4} = \dfrac{3}{9} = \dfrac{1}{3}$，$P(B|-) = \dfrac{2+2}{5+4} = \dfrac{4}{9}$

$P(C|+) = \dfrac{4+2}{5+4} = \dfrac{6}{9} = \dfrac{2}{3}$，$P(C|-) = \dfrac{5+2}{5+4} = \dfrac{7}{9}$

(4) 与(2)同理，此时，

$P_+ = P(A=0|+)\cdot P(B=1|+)\cdot P(C=0|+)\cdot P(+)$

$= (1-\frac{5}{9})\times\frac{1}{3}\times(1-\frac{2}{3})\times\frac{5}{10} \approx 0.0247$

$P_- = P(A=0|-)\cdot P(B=1|-)\cdot P(C=0|-)\cdot P(-)$

$= (1-\frac{4}{9})\cdot\frac{4}{9}\times(1-\frac{7}{9})\times\frac{5}{10} \approx 0.0274$

因为 $P_- > P_+$，所以预测类标号为 $-$.

(5) 使用Laplace估计得到概率更好，因为可以避免条件概率计算结果为0的情况.

8、计算数据点 $x=5.0$ 距各点距离如下表：

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y	−	−	+	+	+	−	−	+	−	−
dis(X,x)	4.5	2.0	0.5	0.4	0.1	0.2	0.3	0.5	2.0	4.5

① 1-最近邻：{+}，结果为+.　② 3-最近邻：{+,−,−}，结果为−.

③ 5-最近邻：{+,−,−,+,+}，结果为+　④ 9-最近邻：{+,−,−,+,+,+,−,−,−}.
分类结果为−