4.1

$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k$ [解析]：

熵是度量样本集合纯度最常用的一种指标，代表一个系统中蕴含多少信息量，信息量越大表明一个系统不确定性就越大，就存在越多的可能性。

假定当前样本集合 D 中第 k 类样本所占的比例为 $p_k(k = 1, 2, \dots, |y|)$ ，则 D 的信息熵为：

$$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k$$

其中，当样本 D 中 $|y|$ 类样本均匀分布时，这时信息熵最大，其值为

$$Ent(D) = -\sum_{k=1}^{|y|} \frac{1}{|y|} log_2 \frac{1}{|y|} = \sum_{k=1}^{|y|} \frac{1}{|y|} log_2 |y| = log_2 |y|$$

此时样本D的纯度越小；

相反，假设样本D中只有一类样本，此时信息熵最小，其值为

$$Ent(D) = -\sum_{k=1}^{|y|} \frac{1}{|y|} log_2 \frac{1}{|y|} = -1 log_2 1 - 0 log_2 0 - \dots - 0 log_2 0 = 0$$

此时样本的纯度最大。

4.2

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

[解析]：假定在样本D中有某个**离散特征** a 有 V 个可能的取值 $(a^1, a^2, \dots, a^V)$，若使用特征 a 来对样本集 D 进行划分，则会产生 V 个分支结点，其中第 v 个分支结点包含了 D 中所有在特征 a 上取值为 a^v 的样本，样本记为 D^v，由于根据离散特征a的每个值划分的 V 个分支结点下的样本数量不一致，对于这 V 个分支结点赋予权重 $\frac{|D^v|}{|D|}$，即样本数越多的分支结点的影响越大，特征 a 对样本集 D 进行划分所获得的“信息增益”为

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

信息增益越大，表示使用特征a来对样本集进行划分所获得的纯度提升越大。

缺点：由于在计算信息增益中倾向于特征值越多的特征进行优先划分，这样假设某个特征值的离散值个数与样本集 D 个数相同（假设为样本编号），虽然用样本编号对样本进行划分，样本纯度提升最高，但是并不具有泛化能力。

4.3

$$Gain - ratio(D, a) = \frac{Gain(D, a)}{IV(a)}$$

[解析]：基于信息增益的缺点，$C4.5$ 算法不直接使用信息增益，而是使用一种叫增益率的方法来选择最优特征进行划分，对于样本集 D 中的离散特征 a ，增益率为

$$Gain-ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$$

其中，

$$IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}log_2\frac{|D^v|}{|D|}$$

IV(a) 是特征 a 的熵。

增益率对特征值较少的特征有一定偏好，因此 $C4.5$ **算法选择特征的方法是先从候选特征中选出信息增益高于平均水平的特征，再从这些特征中选择增益率最高的。**

4.5

$$\begin{aligned}Gini(D)&=\sum_{k=1}^{|y|}\sum_{k\neq k'}p_kp_{k'}\\&=1-\sum_{k=1}^{|y|}p_k^2\end{aligned}$$

[推导]：假定当前样本集合 D 中第 k 类样本所占的比例为 $p_k(k=1,2,\dots,|y|)$，则 D 的**基尼值**为

$$\begin{aligned}Gini(p)&=\sum_{k=1}^{|y|}\sum_{k\neq k'}p_kp_{k'}\\&=\sum_{k=1}^{|y|}p_k(1-p_k)\\&=1-\sum_{k=1}^{|y|}p_k^2\end{aligned}$$

4.7 - 4.8

[解析]：样本集 D 中的**连续特征** a，假设特征 a 有 n 个不同的取值，对其进行大小排序，记为 $\{a^1,a^2,\dots,a^n\}$，根据特征 a 可得到 $n-1$ 个划分点 t，划分点 t 的集合为

$$T_a=\{\frac{a^i+a^{i+1}}{2}|1\leq i\leq n-1\}\qquad(4.7)$$

对于取值集合 T_a 中的每个 t 值计算将特征 a 离散为一个特征值只有两个值，分别是 $\{a>t\}$ 和 $\{a\leq t\}$ 的特征，计算新特征的信息增益，找到信息增益最大的 t 值即为该特征的最优划分点。

$$\begin{aligned}Gain(D,a)&=\max_{t\in T_a}\ Gain(D,a)\\&=\max_{t\in T_a}\ Ent(D)-\sum_{\lambda\in\{-,+\}}\frac{|D_t^\lambda|}{|D|}Ent(D_t^\lambda)\end{aligned}\qquad(4.8)$$

脚注：熵

熵的量度正是能量退化的指标。熵亦被用于计算一个系统中的失序现象，也就是计算该系统混乱的程度。熵是一个描述系统状态的函数，但是经常用熵的参考值和变化量进行分析比较，它在控制论、概率论、数论、天体物理、生命科学等领域都有重要应用，在不同的学科中也有引申出的更为具体的定义，是各领域十分重要的参量。