outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

- 信息熵：用来度量一个属性的信息量。

  假定 S 为训练集，S 的目标属性 C 具有 m 个可能的类标号值，C={C1,C2,⋯,Cm}，假定训练集 S 中，Ci 在所有样本中出现的频率为(i=1,2,3,⋯,m)，则该训练集 S 所包含的信息熵定义为：

$$Entropy(S) = Entropy(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

  熵越小表示样本对目标属性的分布越纯，反之熵越大表示样本对目标属性分布越混乱。

- 解答：令 weather 数据集为 S，其中有 14 个样本,目标属性 play ball 有 2 个值{C1=yes, C2=no}。14 个样本的分布为：
  - 9 个样本的类标号取值为 yes，5 个样本的类标号取值为 No。C1=yes 在所有样本 S 中出现的概率为 9/14，C2=no 在所有样本 S 中出现的概率为 5/14。
  - 因此数据集 S 的熵为：

$$Entropy(S) = Entropy(\frac{9}{14}, \frac{5}{14}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

信息增益

信息增益是划分前样本数据集的不纯程度(熵)和划分后样本数据集的不纯程度(熵)的差值

- 假设划分前样本数据集为 S,并用属性 A 来划分样本集 S，则按属性 A 划分 S 的信息增益 Gain(S,A)为样本集 S 的熵减去按属性 A 划分 S 后的样本子集的熵：

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

按属性 A 划分 S 后的样本子集的熵定义如下：假定属性 A 有 k 个不同的取值，从而将 S 划分为 k 个样本子集{S1,S2,⋯,Sk}，则按属性 A 划分 S 后的样本子集的信息熵为:

$$Entropy_A(S) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} Entropy(S_i)$$

其中|Si|(i,=1,2,⋯k)为样本子集 Si 中包含的样本数，|S|为样本集 S 中包含的样本数。信息增益越大，说明使用属性 A 划分后的样本子集越纯，越有利于分类。

信息增益例题

- 以数据集 weather 为例，设该数据集为 S，假定用属性 wind 来划分 S，求 S 对属性 wind 的信息增益。

- 解答：
  - (1)首先由前例计算得到数据集 S 的熵值为 0.94；
  - (2)属性 wind 有 2 个可能的取值{weak,strong}，它将 S 划分为 2 个子集：{S1,S2},S1 为 wind 属性取值为 weak 的样本子集，共有 8 个样本；S2 为 wind 属性取值为 strong 的样本子集，共有 6 个样本；下面分别计算样本子集 S1 和 S2 的熵。

对样本子集 S1，play ball=yes 的有 6 个样本，play ball=no 的有 2 个样本，则：

$$Entropy(S_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

对样本子集 S2，play ball=yes 的有 3 个样本，play ball=no 的有 3 个样本，则：

$$Entropy(S_2) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

- 利用属性 wind 划分 S 后的熵为：

$$Entropy_{Wind}(S) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} Entropy(S_i) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

$$= \frac{8}{14} Entropy(S_1) + \frac{6}{14} Entropy(S_2) = 0.571 * 0.811 + 0.428 * 1 = 0.891$$

如果一个属性的信息增益量越大，这个属性作为一棵树的根节点就能使这棵树更简洁。

按照信息增益最大的原则选择根节点，子节点的选择重复信息增益计算的步骤。

ID3 优点是理论清晰、方法简单、学习能力较强，但也存在一些缺点：

（1）只能处理分类属性的数据，不能处理连续的数据；

（2）划分过程会由于子集规模过小而造成统计特征不充分而停止；

（3）ID3 算法在选择根节点和各内部节点中的分支属性时，采用信息增益作为评价标准。信息增益的缺点是倾向于选择取值较多的属性（属性里分类多的属性，如分类为多云，下雨，刮风，暴雨，下雪这种长属性），在有些情况下这类属性可能不会提供太多有价值的信息。