3.2 决策树算法的一般过程（C4.5）

（1）树以代表训练样本的单个节点开始；

（2）如果样本都在同一个类中，则这个节点成为树叶结点并标记为该类别；

（3）否则算法使用信息增益率作为启发知识来帮助选择合适的将样本分类的属性，以便将样本集划分为若干子集，

（4）对测试属性的每个已知的离散值创建一个分支，并据此划分样本；

（5）算法使用类似的方法，递归地形成每个划分上的样本决策树：

（6）整个递归过程在下列条件之一成立时停止。

C4.5算法的核心是在决策树各级结点上选择属性时，用信息增益率作为属性的选择标准，以使得在每一个非结点进行测试时，能获得关于被测试记录最大的类别信息。

3.3决策树规则

- 决策树每一条路径都可使用一条产生式规则来解释，整个决策树可以被映射为一组规则。

  Courses≤5
  | Weather = Sunny: Yes (5.0)
  | Weather = Rain: No (3.0/2.0)
  Courses >5: No (5.0)

- **将以上产生的规则翻译为三条产生式规则**

  （1）**IF** Courses ≤5 and Weather = Sunny **THEN** Play = Yes
  正确率：5/5 = 100% 覆盖率：5/7 = 71.4%
  （2）**IF** Courses <≤5 and Weather = Rain **THEN** Play = No
  正确率：3/5 = 60% 覆盖率：3/8 = 37.5%
  （3）**IF** Courses > 5 **THEN**
  正确率：5/5 = 100% 覆盖率：5/8 = 62.5%

简化或淘汰规则

- 例如，若出现如下一条规则：
  - **IF** Courses ≤5 and Weather = Sunny and Temperature = 20~30 **THEN** Play = Yes
  - 正确率：2/2 = 100% 覆盖率：2/7 = 28.6%
  - 可简化为
    - **IF** Courses ≤5 and Weather = Sunny **THEN** Play = Yes
    - 正确率：5/5 = 100% 覆盖率：5/7 = 71.4%