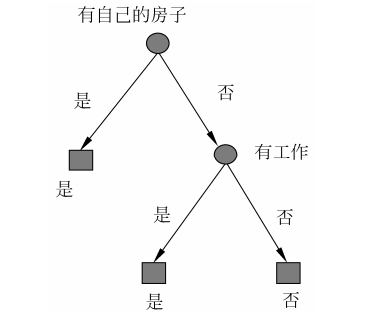
**《决策树模型三》**

**决策树生成算法-ID3算法**

ID3 算法的核心是在决策树各个结点上应用信息增益准则选择特征，进行决策点的划分，从而构建决策树。

**具体方法是**：从根结点开始，对结点计算所有可能的特征的信息增益，选择信息增益最大的特征作为结点的特征，由该特征的不同取值建立子结点；再对子结点递归地调用以上方法，构建决策树；直到所有特征的信息增益均很小或没有特征可以选择为止．可以看出这种构建的方式是一种递归方式来进行工作的。



基于ID3算法对表5.1的数据构建的决策树

ID3优缺点：

优点：简单，清晰

**缺点：**（1）ID3基于信息增益，信息增益的缺点是倾向于选择取值较多的属性，在某些情况下，这类属性可能不会提供太多有价值的信息。

1. 不能处理具有连续值的属性，也不能处理具有缺失值的属性。
2. 没有剪枝等操作，存在着一定的过拟合的风险。

**ID3算法对数据的要求：**  
1) 所有属性必须为离散量；  
2) 所有的训练例的所有属性必须有一个明确的值；  
3) 相同的因素必须得到相同的结论且训练例必须唯一

**为什么倾向特征选项较多的特征**

原因：信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大。