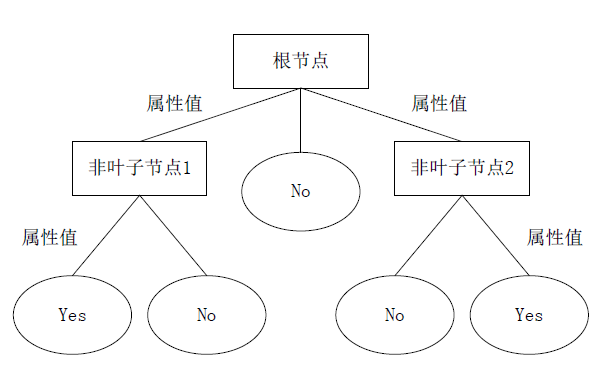
决策树与随机森林

决策树原理



决策树通过把样本实例从根节点排列到某个叶子节点来对其进行分类。树上的每个非叶子节点代表对一个属性取值的测试，其分支就代表测试的每个结果；而树上的每个叶子节点均代表一个分类的类别，树的最高层节点是根节点。

决策树采用自顶向下的递归方式，从树的根节点开始，在它的内部节点上进行属性值的测试比较。然后按照给定实例的属性值确定对应的分支，最后在决策树的叶子节点得到结论。这个过程在以新的节点为根的子树上重复。

决策树的优点与缺点

优点：

决策树容易理解和实现。

对于决策树，数据的准备往往是比较简单或者是不必要的。其它技术往往要求先把数据归一化，比如去掉多余的或者空白的属性。

能够同时处理数据型和常规型属性。其它的技术往往要求数据属性的单一。

是一个白盒模型。如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。

缺点

对于各类别样本数量不一致的数据，在决策树当中信息增益的结果偏向于那些具有更多数值的特征。

决策树内部节点的判别具有明确性，这种明确性可能会带来误导。

随机森林

随机森林算法基于Bootstrap方法重采样，产生多个训练集。不同的是，随机森林算法在构建决策树的时候，采用了随机选取分裂属性集的方法。